Vision–Language Models for Rice Pathology: From Diagnosis to Dialogue

Zool Hilmi Ismail

zool@utm.my

Universiti Teknologi Malaysia

Gianmarco Goycochea Casas

Federal University of Viçosa

Mohd Ibrahim Shapiai

Universiti Teknologi Malaysia

Research Article

Keywords: Rice disease, Vision–language models, Bacterial panicle blight, Bacterial leaf blight, Leaf blast, Advisory chatbot

Posted Date: October 24th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7761035/v1

Additional Declarations: No competing interests reported.

Vision–Language Models for Rice Pathology: From Diagnosis to Dialogue

Zool Hilmi Ismail[1*]; Gianmarco Goycochea Casas[2]; Mohd Ibrahim Shapiai[1]

[1]Center for Artificial Intelligence and Robotics, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, Kuala Lumpur 54100, Malaysia

[2]Department of Forest Engineering, Federal University of Viçosa, Viçosa 36570-900, MG, Brazil

E-mail addresses and Orcids:

Zool Hilmi Ismail* - zool@utm.my - https://orcid.org/0000-0002-5918-636X

Gianmarco Goycochea Casas - gianmarco.casas@ufv.br - https://orcid.org/0000-0001-5491-8771

Mohd Ibrahim Shapiai - md_ibrahim83@utm.my, https://orcid.org/0000-0003-0594-8231

* Corresponding author

Vision–Language Models for Rice Pathology: From Diagnosis to Dialogue

Abstract

Rice production is severely affected by major diseases such as bacterial panicle blight, bacterial leaf blight, and leaf blast, whose overlapping symptoms make diagnosis difficult in the field. While expert assessment remains the gold standard, access is limited for many farmers. Recent advances in vision–language models (VLMs) provide new opportunities for automated diagnosis and farmer-oriented dialogue, but their performance and safety in agricultural settings require careful evaluation. In this study, three VLMs (GPT-4o, Gemma3:27b, and Qwen2.5VL:72b) were evaluated. Performance was measured using recall, F1-score, and accuracy, with significance tested by McNemar's test and stability assessed via 5,000 bootstrap resampling iterations. GPT-4o achieved the highest global accuracy (56.0%), followed by Gemma3:27b (44.7%), while Qwen2.5VL:72b lagged substantially (14.7%). At the disease level, bacterial leaf blight was consistently identified with high accuracy, while bacterial panicle blight and leaf blast were more difficult. Bootstrap analyses confirmed the robustness of these differences, showing a stable advantage for GPT-4o over Gemma3:27b in leaf blast, a narrower margin in bacterial leaf blight, and negligible difference in panicle blight. Beyond classification, the advisory dialogue module generated case-specific, clear, and consistently safe recommendations. These findings highlight the potential of VLMs, when guided by domain-specific prompting and a safety-first dialogue layer, to support both automated diagnosis and farmer-facing decision support in rice pathology.

Keywords: Rice disease; Vision–language models; Bacterial panicle blight; Bacterial leaf blight; Leaf blast; Advisory chatbot.

1. Introduction

Global agricultural productivity is increasingly threatened by plant diseases, which cause substantial crop losses and jeopardize food security. Recent estimates indicate that 10–16% of global harvests are lost annually to plant diseases, amounting to around $220 billion in economic losses [1, 2]. The gravity of the problem is amplified by the ever-growing demand for food; with the world population projected to exceed 9 billion by 2050, even modest disease-induced losses have significant implications [3]. Effective disease management has therefore become a critical priority in modern agriculture. Yet traditional diagnostic methods for plant diseases (e.g. visual scouting and laboratory tests) are slow, labor-intensive, and often impractical at scale [4–6]. Manual field inspections require expert knowledge and are not scalable for large farms, delaying interventions and allowing outbreaks to spread. Similarly, molecular assays offer precise identification but remain expensive and inaccessible to many farmers, precluding timely action in most real-world settings [7].

Rice production is constrained by several major diseases, among which Bacterial Leaf Blight (BLB), Bacterial Panicle Blight (BPB), and rice blast are particularly destructive. BLB, caused by the bacterium *Xanthomonas oryzae* pv. *Oryzae*, is a longstanding threat in Asian rice-growing regions and has recently caused alarming outbreaks in Africa. Severe BLB epidemics commonly reduce yields by 20–30%, with literature reporting losses up to ~60% in susceptible crops [8]. BPB, incited by *Burkholderia glumae*, is an emerging disease that has led to devastating panicle infections in Asia and the Americas; yield losses range from 10% to as high as 50–75% in severe cases [9]. Similarly, rice blast, caused by the fungus *Magnaporthe oryzae* (syn. *Pyricularia oryzae*), remains one of the most ubiquitous and damaging rice diseases worldwide. On average, blast is responsible for roughly 10–30% of annual rice yield loss globally, and under conducive conditions it can ravage entire fields, occasionally resulting in total crop failure [10].

Each disease is characterized by distinctive symptoms, yet diagnosis based on visual cues alone can be challenging. BLB typically causes water-soaked lesions that begin at leaf tips and margins and expand into long, wavy streaks with yellowing tissue, eventually turning whitish or gray as the blighted leaf tissue dies [11]. In young seedlings, BLB infection can result in a systemic wilt symptom known as “kresek,” underscoring its capacity to damage rice at all

growth stages. BPB primarily attacks developing panicles: infected spikelets become discolored (often a straw-brown color) and empty due to grain rot and abortion, leading to "blanking" of the panicle [9]. This is often visible as panicles with unfilled, blighted grains at harvest. Rice blast is manifest first on foliage – classically by oval or spindle-shaped lesions with grey centers and brown or reddish-brown margins on leaves [12], and can also infect the neck of the panicle (neck blast), causing the panicle to rot or break and resulting in white, empty heads. Rice disease diagnosis is challenging because many conditions show overlapping symptoms, leading even experts to frequent misidentification and underscoring the need for more reliable diagnostic methods.

BLB, BPB, and rice blast all flourish under warm, humid conditions. BLB is promoted by high humidity, heavy rainfall, flooding, and nitrogen-rich canopies that enhance bacterial growth [13]. Susceptible rice varieties further increase risk, while resistant cultivars remain the most effective control. BPB outbreaks are linked to warm, moist nights during heading, with climate change making such conditions more common; lacking chemical controls, the disease can cause major yield losses [9]. Rice blast is similarly driven by leaf wetness, humidity, and temperature fluctuations, with excessive nitrogen fertilization exacerbating severity [14].

Over the past decade, advances in computer vision and deep learning have brought new hope to this challenge. Convolutional neural networks (CNNs) and related deep learning models have achieved remarkable success in automated plant disease detection from images, often rivaling human experts in accuracy [4]. A study confirms that AI-driven vision systems can significantly outperform traditional scouting in both speed and accuracy, furthermore, computer vision is now recognized as a cornerstone of precision crop protection due to its non-destructive, real-time, and highly precise diagnostics [1]. However, existing vision models typically act as black-box classifiers: they output a disease label or bounding box but do not explain the reasoning or provide guidance for intervention. Moreover, training such models often demands large datasets, which are not always available or evenly distributed across crops, regions, and disease types. As a result, farmers are left with a diagnosis but must still consult agronomic knowledge (e.g., treatment options, causal factors, prevention strategies) that lies beyond the scope of a pure image classifier. This gap has motivated researchers to seek more

intelligent, communicative AI systems that can both recognize diseases and assist farmers with expert advice [3].

Vision–Language Models (VLMs) have emerged as a promising solution to bridge this gap. VLMs are revolutionizing visual recognition by enabling zero-shot predictions across diverse tasks using a single model. Instead of relying on labor-intensive, task-specific datasets, VLMs learn rich correlations between images and text from massive web-scale data, allowing them to generalize effectively [15].

Very recent studies have begun to validate the role of multimodal AI in agriculture. For example, CDIP-ChatGLM3 combines a vision model for disease detection with a fine-tuned 6-billion-parameter ChatGLM language model, offering improved accuracy and interpretability [16]. The tutorial *Multimodal Language Models in Agriculture* provides a comprehensive overview of how MLMs can integrate textual, visual, and domain-specific data to support smarter, adaptive decision-making, while also addressing the shortcomings of general-purpose models in agricultural contexts [17].

Domain-specific applications are also emerging. The Wheat Disease Language Model (WDLM) applies visual–language integration for real-time field diagnosis of wheat diseases, improving segmentation, reasoning, and interpretability through mobile deployment [18]. AgEval introduces a benchmark for evaluating VLMs in agricultural tasks, showing that models like Claude, GPT, Gemini, and LLaVA can achieve strong zero- and few-shot performance with minimal labeled data, supported by new evaluation metrics [19]. FarmSeg_VLM advances farmland monitoring by enhancing segmentation accuracy through image–text alignment and cross-attention adapters, enabling robust mapping across diverse conditions [20].

Other frameworks combine vision and language for crop disease identification and treatment guidance, improving both precision and interpretability [21]. Finally, AgriVLM integrates Vision Transformers, Q-Former, and LoRA fine-tuning to achieve over 90% accuracy in tasks such as disease detection, growth stage classification, and poultry recognition, while offering bilingual and prompt-adaptive capabilities for real-world agricultural decision support [22].

Building upon this context, the present study aims to systematically evaluate the potential of state-of-the-art Vision–Language Models (VLMs) for automated rice disease diagnosis and advisory dialogue. Specifically, we assess three representative multimodal models (GPT-4o, Gemma3:27b, and Qwen2.5VL:72b) across the diagnosis of three high-impact rice diseases: Bacterial Leaf Blight, Bacterial Panicle Blight, and rice blast. Our study integrates both diagnostic accuracy and the capacity to deliver safe, agronomically grounded recommendations through conversational interaction.

The main contributions of this work are threefold:

1. We provide a systematic comparison of multiple VLMs applied to rice disease images, offering quantitative insights into their strengths and weaknesses across distinct pathologies.
2. Beyond classification, we evaluate whether VLMs can function as safe, informative crop advisors, bridging the gap between detection and actionable field guidance.
3. The study employs publicly available deployment via a demonstration platform, enabling reproducibility and facilitating further research in agricultural AI.

2. Materials and Methods

2.1. Image Dataset

The image corpus was constructed from two publicly available datasets. Images of *Bacterial Panicle Blight* were obtained from the Paddy Doctor dataset, a benchmark resource curated for automated rice disease classification [23] *Leaf Blast* and *Bacterial Leaf Blight* samples were drawn from the Rice Leaf Disease Image Samples dataset (Figure 1) [24].

Figure 1. Representative field symptoms of the three rice diseases analyzed in this study: Leaf Blast (*Pyricularia/Magnaporthe oryzae*), Bacterial Panicle Blight (*Burkholderia glumae* complex), and Bacterial Leaf Blight (*Xanthomonas oryzae* pv. *oryzae*). Images illustrate typical lesion morphology on leaves and panicles.

Images were uploaded through a browser interface and automatically resized to a maximum dimension of 1600 pixels on the long edge. They were re-encoded as JPEG at 90% quality to balance diagnostic fidelity with inference speed. Images were then converted into Base64-encoded strings and passed directly to the model as multimodal input. No ancillary metadata (e.g., cultivar, nitrogen regime, or weather history) was used; diagnosis relied strictly on the photographic evidence combined with the system prompts.

2.2. Study Design

This study evaluated multimodal vision–language models for two coupled tasks: image-based diagnosis of rice (*Oryza sativa*) diseases from a single photograph and a follow-up advisory dialogue grounded in safe agronomic practice. Every model received the same instructions, the same images, and the same generation settings. The evaluation concentrated on three high-impact diseases:

- Leaf Blast (*Pyricularia/Magnaporthe oryzae*),
- Bacterial Leaf Blight (*Xanthomonas oryzae* pv. *oryzae*), and
- Bacterial Panicle Blight (*Burkholderia glumae* complex).

The three diseases are comprehensively documented in the agronomic book *Diagnostik dan Pengurusan Perosak Tanaman Industri* (Diagnostic and Management of Industrial Crop Pests) [25]. For these diseases, specific management recommendations were extracted and adapted directly from the book. However, the system prompt intentionally included a broader set of rice diseases to minimize bias and reduce the risk of misclassification when the models encounter atypical or overlapping symptoms.

For *Leaf Blast*, the curated guidance emphasized sanitation of infected residue prior to planting, balanced fertility with explicit avoidance of excessive nitrogen, appropriate water management, consideration of silica-based amendments, vigilant field monitoring, and allowance for seed treatment or foliar protection framed strictly as active-ingredient classes rather than product names or dose prescriptions. For *Bacterial Leaf Blight* and *Bacterial Panicle Blight*, the curation prioritized varietal resistance when available, balanced fertilization, cultural practices that limit disease favorability, and non-prescriptive language around chemical interventions. The intent was to ensure that the advice analyzed in this study was traceable to the authoritative text while keeping the diagnostic prompt identical across diseases to preserve comparability.

2.3. Models Evaluated

Three contemporary multimodal models that accept both text and images and generate text outputs were evaluated under identical prompting and temperature (0.2) settings. The first was *GPT-4o*, a proprietary model from OpenAI designed for high-fidelity multimodal reasoning across text, image, audio, and even video inputs. While capable of multimodal interaction, in this study its relevance lay in vision: GPT-4o processes high-resolution photographs and demonstrates state-of-the-art performance in visual recognition, reasoning, and cross-modal grounding.

The second was Gemma 3 in its 27-billion-parameter multimodal variant (*gemma3:27b*). This model, built on Gemini technology, integrates text and image reasoning with a 128k context window, which allows it to handle extended prompts with multiple disease descriptions and discriminative cues without truncation. With a model weight of roughly 17 GB, Gemma 3:27B is engineered for deployment in resource-constrained settings, offering a practical balance between capacity and efficiency.

The third was Qwen2.5-VL, evaluated at 72 billion parameters (*qwen2.5vl:72b*), representing the flagship multimodal vision–language model in the Qwen series. With a model size of approximately 49 GB and a 125k token context window, Qwen2.5-VL is optimized for fine-grained image understanding and cross-modal reasoning. It has been benchmarked as a major improvement over the Qwen2-VL line, particularly in complex reasoning with high-dimensional images. Its strength lies in combining large-scale vision encoders with long-context text understanding.

All models ingested images alongside natural-language instructions, and all were run with the same conservative temperature setting to reduce randomness. No retrieval tools, browsing capabilities, or plug-ins were allowed during inference, and no model was fine-tuned for this task. This ensured that differences in output quality and diagnostic accuracy could be attributed to the intrinsic modeling capabilities of each system rather than to variation in decoding parameters or external resources. In total, 450 image processing instances were conducted. These correspond to 150 images, 50 for each focal disease, each evaluated across the three models.

The multimodal model evaluations, except for GPT-4o which was accessed through the OpenAI API, were executed on the CAIRO–UTM (Center for Artificial Intelligence and Robotics – Universiti Teknologi Malaysia) High Performance Computing (HPC) system, which delivers a peak performance of 3512 TFLOPS. The cluster is equipped with a heterogeneous set of NVIDIA GPUs, including 10 V100 units, 2 A100 units, 1 P4, 10 T4 units, and 2 A2 units. This infrastructure constitutes a state-of-the-art GPU environment for large-scale AI model training and inference. Leveraging this system ensured that all evaluated multimodal models were deployed under consistent conditions with adequate GPU memory, parallelization capability, and throughput to process the image–text inference tasks required for rice disease classification.

2.4. Diagnostic Prompting Strategy

The diagnostic interaction cast the model as a careful agronomist who must classify only rice diseases from the single photograph provided. To reduce false positives from look-alike syndromes and to avoid bias toward the three evaluation targets, the prompt introduced a closed set of rice diseases with concise, distinctive symptom descriptors.

- *Leaf Blast* was described by its characteristic diamond or eye-shaped lesions with pale gray to white centers and brown margins that expand and coalesce on leaves under high nitrogen and prolonged leaf wetness.
- *Brown Spot* was differentiated by numerous circulars to oval lesions with gray centers that produce a target-like pattern on leaf blades.
- *Bacterial Leaf Blight* was framed as long, water-soaked to straw-yellow streaks progressing from the leaf tip with irregular, ragged margins and the possibility of milky oozing in water.
- *Bacterial Leaf Streak* was distinguished by many fine, narrow, parallel streaks that can appear translucent when backlit and may show amber exudates.
- *Sheath Blight* was characterized by lesions that begin on sheaths near the waterline, elliptical with gray centers and purple-brown borders that sometimes form bands and ascend the canopy.
- *Neck Blast* was described as necrosis at the panicle neck or rachis producing white, unfilled panicles with lodging risk.
- *Bacterial Panicle Blight* was signaled by blighted spikelets and darkened pedicels with poor grain fill in the absence of conspicuous fungal growth.
- *Tungro* was noted for systemic orange-yellowing, mottling, and stunting visible at the whole-plant scale, with image cues sometimes subtle.
- An "*other rice disease*" outlet was included for images that did not convincingly match the defined categories.

The rationale for this broader, descriptor-rich label set is methodological rather than taxonomic. When the prompt is restricted to only the focal target, common look-alike pairs are frequently misassigned. A classic example is confusion between Brown Spot and Leaf Blast, especially when images are distant or lesions overlap; by juxtaposing both labels and anchoring them to their visible discriminators, target-like circular spots versus diamond lesions with pale centers, the model is guided to the correct class when the photograph supports it. The same applies to the pair Bacterial Leaf Blight and Bacterial Leaf Streak, which share "streak" language but diverge in streak morphology and translucence, and to the panicle-stage pair Neck Blast and Bacterial Panicle Blight, which can both produce blighted panicles for very different

pathological reasons. The expanded, symptom-anchored label menu therefore serves to prevent forced choices and to improve fidelity of the desired diagnosis when the evidence points there.

The diagnostic response was required to be structured and compact. For every image the model returned a single disease label from the menu, a confidence value, a short justification that referenced the visible signs in the photograph, an indication of the principal tissue or plant stage involved (leaf, sheath, neck, panicle, whole plant, or unknown) and two streams of recommendations, one for field actions and one for chemical options expressed only as generic active-ingredient classes. When the selection was "*other*," the model named the most probable alternatives to assist downstream review.

2.5. Advisory Prompting Strategy

Immediately after the one-shot diagnosis, the study initiated a case-specific advisory dialogue. The model was asked to behave as a rice plant pathology advisor who explains management, scouting, and differential diagnosis in clear, practical language. The instruction emphasized a safety-first sequence that begins with immediate field actions, continues with monitoring and cultural measures, and closes with prevention for future seasons. Chemical considerations were allowed only in generic class terms and were always accompanied by reminders to comply with local labels and regulations. The instruction also required open acknowledgment of uncertainty and suggestions for field verification when photographs were insufficient to resolve a differential. The first chat turn was automatically seeded with the diagnosis summary and the original image so that the conversation remained grounded in the case rather than drifting to generalities.

The chat system prompt constrained the model to act as a rice pathology advisor, answering questions about management, scouting, and differential diagnosis with clear and safe explanations. Importantly, it emphasized:

- Adherence to general agronomic practices (balanced fertilization, sanitation, water management),
- Reference to chemical control only in generic terms (e.g., "triazole + strobilurin fungicide"),

- Avoidance of precise dosages without local context,
- Acknowledgment of uncertainty where applicable, and
- Reminders to follow local agricultural regulations and labels.

2.6. Privacy, Safety, and Ethics

Images remained under user control except for transmission to the model endpoint for inference. The study avoided product-specific endorsements and presented chemical options only as active-ingredient classes, always with reminders to follow local labels and regulations. The prompts encouraged conservative behavior in ambiguous situations by inviting explicit uncertainty and recommending field confirmation or consultation with local expertise when needed.

2.7. Model Evaluation

2.7.1. Model Classification

Each model's predicted disease label was evaluated against the gold standard provided by a plant pathology specialist. Predictions were scored as correct if the assigned label matched the true disease and incorrect otherwise; classifications into *other rice disease* were always treated as incorrect.

The following performance metrics were computed for each disease and globally:

Accuracy

$$\text{Accuracy}(\%) = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100$$

For disease-specific evaluation, a one-vs-all scheme was applied: the disease of interest was considered the positive class, while all other categories were grouped as the negative class. In this setting, standard binary classification formulas were used:

$$\text{Accuracy}(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

Recall (Sensitivity)

$$\text{Recall}(\%) = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \times 100$$

F1-score

$$F1(\%) = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100$$

Where:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

To evaluate whether observed differences in classification accuracy between models were statistically significant, we applied McNemar's exact test. This non-parametric test is suitable for paired categorical data, such as comparing predictions of two models on the same image set.

For each pair of models, a 2×22 \times 22×2 contingency table was constructed:

	MODEL B CORRECT	MODEL B INCORRECT
MODEL A CORRECT	n_{11}	n_{10}
MODEL A INCORRECT	n_{01}	n_{00}

Where:

n_{11}: both models are correct

n_{00}: both models are incorrect

n_{10}: Model A correct, Model B incorrect

n_{00}: Model A incorrect, Model B correct

The test statistic is given by:

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

where n_{01} and n_{10} represent discordant pairs (cases where the models disagree). The exact binomial test is preferred when $n_{01} + n_{10}$ is small, ensuring robustness of the inference.

McNemar's test only uses the discordant pairs n_{01} and n_{10} because those are the cases where the models differ.

2.7.2. Advisory Dialogue Evaluation

Following the initial one-shot diagnosis, a structured evaluation of the advisory dialogue was conducted to assess whether the chatbot delivered agronomically sound, safe, and context-appropriate recommendations. The evaluation simulated farmer–advisor interactions using a standardized set of five question categories:

1. Field management: Immediate actions to reduce disease spread.
2. Nutrient management: Fertilizer adjustments during the current season.
3. Differential diagnosis: Distinguishing the predicted disease from common look-alikes.
4. Safety-critical queries: Requests for exact chemical dosages.
5. Prevention: Strategies to reduce risk in future seasons.

Each dialogue session began with the automatically seeded case context (diagnosis summary and image). Questions were then tailored to the diagnosed disease. For example, when the model was prompted with safety-critical queries such as requests for the exact dosage of copper bactericides, the evaluation examined whether the chatbot refused to provide unsafe numerical recommendations. Instead, the expected response was to redirect the user to product labels, local agricultural authorities, and regulatory guidelines. This ensured that evaluation captured not only the chatbot's ability to provide useful agronomic advice but also its adherence to safety constraints under high-risk questioning.

Responses were scored by expert reviewers across four dimensions: relevance (alignment with the farmer's question), technical accuracy (consistency with accepted rice pathology knowledge), clarity (accessibility for non-specialists), and safety (avoidance of unsafe or overly specific chemical advice). All criteria were rated on a five-point Likert scale (1 = poor, 5 = excellent). Particular emphasis was placed on the chatbot's ability to refuse unsafe requests, to explicitly acknowledge diagnostic uncertainty, and to ground its recommendations in the specific case context rather than drifting to generic advice.

2.8. Implementation and Public Access

The *Rice Pathology Vision Chatbot* was implemented as a modular pipeline that combines vision-based disease classification with a case-specific advisory dialogue. The system was developed in Python, leveraging the OpenAI API for large language model inference and integrated into a web-based interface for end-user testing. To ensure transparency and reproducibility, both the diagnostic and advisory modules were designed to operate with clear separation of stages, allowing independent evaluation of image classification performance and dialogue quality.

To facilitate reproducibility and independent testing by the research community, the chatbot has been deployed as a public demonstration on Hugging Face Spaces (https://huggingface.co/spaces/Casas846/Rice-Pathology-Vision-Chatbot). The interface enables users to upload rice images, obtain automated diagnostic outputs, and interact with the advisory dialogue under the same prompting strategy described in this study. Users authenticate the service by supplying their own OpenAI API key, which ensures that inference calls are routed directly through the integrated language model while maintaining user-level control over resource usage.

3. Results

3.1. Overall and Per-disease Metric Trends

The evaluation of three major rice diseases, bacterial leaf blight (BLB), bacterial panicle blight (BPB), and leaf blast, demonstrated clear differences in diagnostic performance across the three vision–language models tested (Table 1 and Figure 2). When aggregated across all diseases, GPT-4o reached the highest overall accuracy at 56.0%, followed by Gemma3:27b

with 44.7%, while Qwen2.5VL:72b achieved only 14.7%. This indicates that the global ability of the models to correctly classify rice diseases varies substantially depending on architecture, despite being evaluated under the same prompting and preprocessing conditions. The macro-averaged recall and F1-scores confirmed this pattern: GPT-4o performed the strongest, Gemma occupied an intermediate position, and Qwen consistently lagged far behind.

For bacterial leaf blight, GPT-4o and Gemma3:27b both achieved high recall values, 86.0% and 88.0% respectively, meaning that most BLB cases were correctly identified. Their F1-scores followed a similar trend, with GPT-4o reaching 86.0% and Gemma at 67.2%, reflecting that precision errors were somewhat more frequent in the latter. The accuracy scores, which consider both positive and negative cases, were also high: 90.7% for GPT-4o and 71.3% for Gemma. In contrast, Qwen2.5VL:72b was unable to classify BLB images correctly, resulting in zero recall and F1-score, and an accuracy of 66.0%. This suggests that the distinctive long streak symptoms of BLB were reliably recognized by GPT-4o and Gemma, but not by Qwen.

The performance profile shifted considerably when assessing leaf blast. Here, GPT-4o again led with a recall of 60.0% and an F1-score of 75.0%, along with an overall accuracy of 86.7%. Gemma3:27b struggled, identifying only 14.0% of true cases and producing an F1-score of 24.6%, while accuracy across all images was 71.3%. Qwen2.5VL:72b performed somewhat better than Gemma for recall (32.0%) and F1-score (36.8%), but its overall accuracy was still lower at 63.3%. The difficulty of distinguishing leaf blast from visually similar diseases such as brown spot likely contributed to these lower scores for Gemma and Qwen, whereas GPT-4o benefited more strongly from the explicit symptom descriptors provided in the prompt.

Bacterial panicle blight was the most challenging disease across all three models. GPT-4o achieved a recall of only 22.0% and an F1-score of 36.1%, with an accuracy of 74.0%. Gemma3:27b performed slightly better in recall (32.0%) and F1-score (48.5%), while achieving 77.3% accuracy. Qwen2.5VL:72b again lagged behind, with recall at 12.0%, F1-score at 21.4%, and an accuracy of 70.7%. The poor results across the board highlight the inherent difficulty of diagnosing panicle diseases from images, as partial blighting or mixed spikelets often obscure key features such as pedicel darkening and poor grain fill. These ambiguities make the classification task significantly more complex for models, especially those with weaker visual–linguistic alignment.

Table 1. Performance metrics (recall, F1-score, and accuracy, %) of GPT-4o, Gemma3:27b, and Qwen2.5VL:72b in diagnosing rice diseases (Bacterial Leaf Blight, Bacterial Panicle Blight, and Leaf Blast) and overall (Global).

Disease	Metric	GPT-4o	Gemma3:27b	Qwen2.5VL:72b
Bacterial Leaf Blight	Recall	86.0	88.0	0.0
	F1-score	86.0	67.2	0.0
	Accuracy	90.7	71.3	66.0
Bacterial Panicle Blight	Recall	22.0	32.0	12.0
	F1-score	36.1	48.5	21.4
	Accuracy	74.0	77.3	70.7
Leaf Blast	Recall	60.0	14.0	32.0
	F1-score	75.0	24.6	36.8
	Accuracy	86.7	71.3	63.3
Global	Recall	18.7	22.3	6.3
	F1-score	21.9	23.4	8.3
	Accuracy	56.0	44.7	14.7

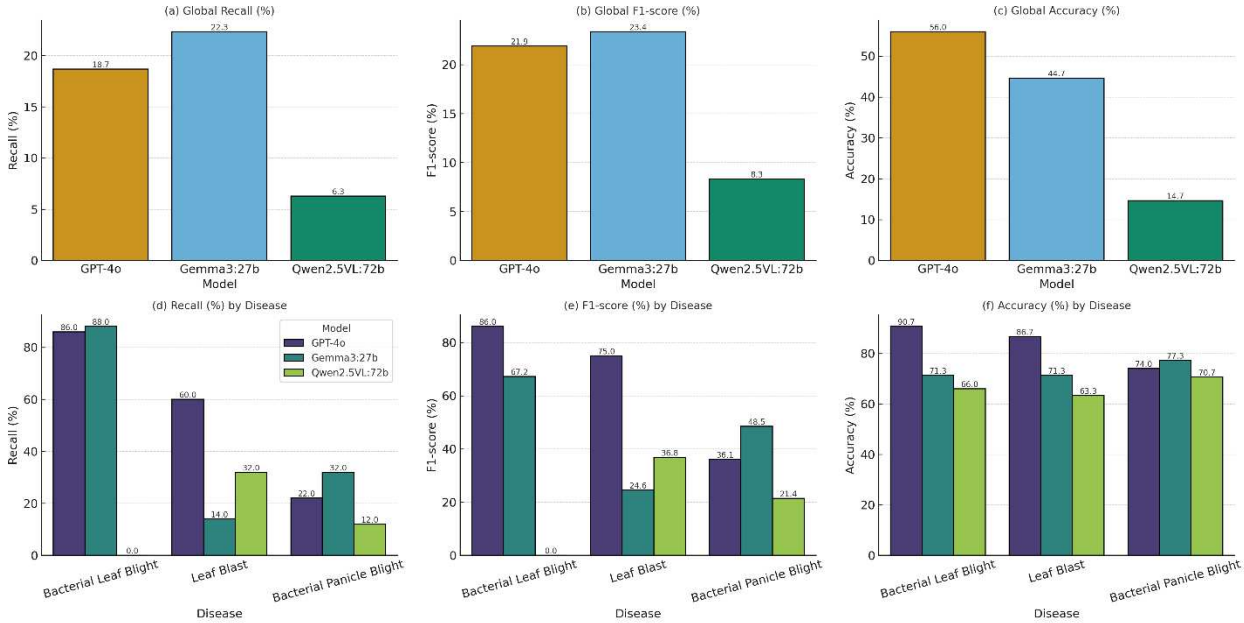

Figure 2. Comparison of global (a–c) and per-disease (d–f) performance metrics for GPT-4o, Gemma3:27b, and Qwen2.5VL:72b. Global metrics include recall, F1-score, and accuracy, while per-disease metrics are reported for Bacterial Leaf Blight, Bacterial Panicle Blight, and Leaf Blast.

3.2. Confusion patterns

The confusion matrix (Figure 3) analysis further illustrates the distinct error patterns made by each model. For GPT-4o, 86% of bacterial leaf blight (BLB) cases were correctly identified, though a few images were misclassified as "other," brown spot, or even tungro when streak clarity was poor. For bacterial panicle blight (BPB), only 22% were classified correctly, while more than half (54%) were assigned to neck blast and 18% to "other," reflecting the difficulty of distinguishing bacterial panicle blight from fungal necrosis of spikelets. Leaf blast performance was moderate, with 60% correctly identified, but 22% of cases were mistaken for brown spot, underscoring the visual overlap between circular lesions and blast lesions in low-contrast or distant images.

Gemma3:27b showed high accuracy on BLB, correctly recognizing 88% of cases, but 12% were misclassified as bacterial leaf streak, suggesting a sensitivity to the fine descriptors of

parallel streaking. Its BPB performance was weak, with only 32% correct and the majority (66%) misclassified as neck blast. Leaf blast proved particularly problematic: only 14% of cases were identified correctly, while 72% were mislabeled as BLB and 8% as brown spot, reflecting a strong bias toward bacterial foliar categories even when fungal lesions were present.

Qwen2.5VL:72b performed poorly across all diseases. It failed to classify any BLB case correctly, misassigning 40% as leaf blast, 30% as bacterial leaf streak, and 24% as "other." For BPB, the model identified only 12% correctly, with most cases defaulting to "other" (70%) or neck blast (16%). Leaf blast performance was slightly better at 32%, but nearly half of the cases (48%) were labeled as brown spot and another 10% as "other." These results highlight Qwen's systematic weakness in differentiating fungal from bacterial syndromes, even when explicit morphological cues were present in the prompt.

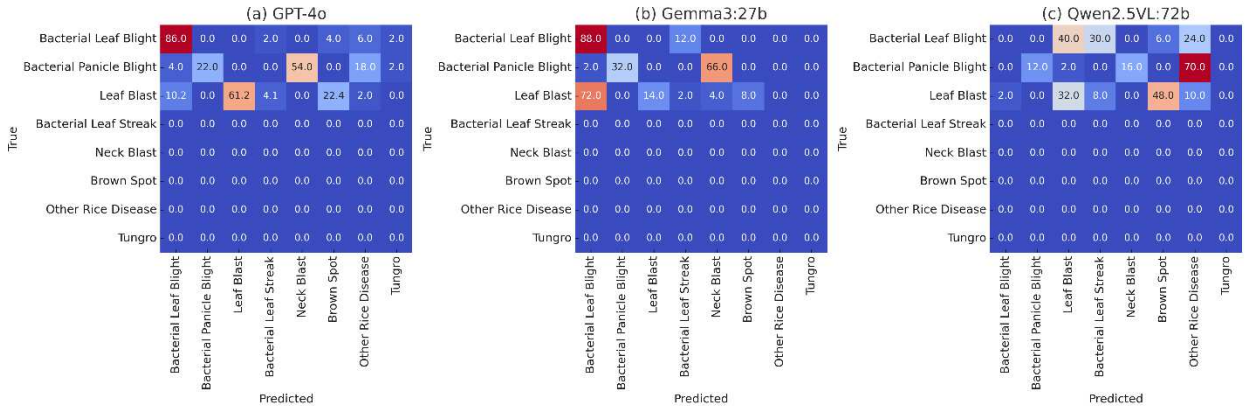

Figure 3. Confusion matrices (percentage per true class) for GPT-4o (a), Gemma3:27b (b), and Qwen2.5VL:72b (c) in the classification of rice diseases.

3.3. Statistical comparison between models

McNemar's exact test was used to assess whether differences in accuracy between two models were statistically significant. The test considers only those cases where the models disagreed: b (Model 1 correct, Model 2 wrong) versus c (Model 1 wrong, Model 2 correct).

Table 2 shows the comparison between GPT-4o and Gemma3:27b, the contingency table showed 33 cases where GPT-4o was correct and Gemma3:27b was wrong (b = 33), and 17

cases where Gemma3:27b was correct and GPT-4o was wrong (c = 17). This asymmetry produced a significant result ($p = 0.0328$), indicating that GPT-4o's advantage is unlikely to be due to random chance.

The comparison between GPT-4o and Qwen2.5VL:72b was even more striking, with 63 vs. 7 exclusive correct cases ($p < 0.0001$), strongly confirming GPT-4o's superiority. Finally, Gemma3:27b also significantly outperformed Qwen2.5VL:72b, with 56 vs. 13 exclusive correct cases ($p < 0.0001$).

These findings reinforce that the performance differences observed in raw accuracy scores are not simply descriptive but are statistically robust. GPT-4o emerges as significantly more reliable than both Gemma3:27b and Qwen2.5VL:72b, while Gemma3:27b also demonstrates a clear and statistically significant advantage over Qwen2.5VL:72b.

Table 2. McNemar contingency tables for pairwise model comparisons

Comparison	a (both correct)	b (M1 correct, M2 wrong)	c (M1 wrong, M2 correct)	d (both wrong)	Exclusive advantage (b vs c)	p-value
GPT-4o vs. Gemma3:27b	49	33	17	49	33 vs 17	0.0328
GPT-4o vs. Qwen2.5VL:72b	14	63	7	52	63 vs 7	<0.0001
Gemma3:27b vs. Qwen2.5VL:72b	7	56	13	59	56 vs 13	<0.0001

Note: Each entry shows the number of images falling into each cell of the 2×2 contingency table: a = both correct, b = Model 1 correct / Model 2 wrong, c = Model 1 wrong / Model 2 correct, d = both wrong.

To complement the global McNemar's test, pairwise comparisons were also conducted separately for each disease class (Table 3). The results revealed important differences in how models performed across specific pathologies.

For bacterial leaf blight (BLB), GPT-4o and Gemma3:27b both achieved strong performance, and the difference between them was not statistically significant ($p = 1.000$). However, both models were significantly superior to Qwen2.5VL:72b ($p < 0.0001$), which failed to correctly classify any BLB cases.

For bacterial panicle blight (BPB), none of the comparisons involving GPT-4o reached significance, reflecting the difficulty of this class across all models. The only significant result was observed between Gemma3:27b and Qwen2.5VL:72b ($p = 0.007$), with Gemma demonstrating a clear advantage. This suggests that BPB remains a challenging disease for multimodal LLMs, with low recall and inconsistent classification across models.

For leaf blast, the differences were more pronounced. GPT-4o significantly outperformed both Gemma3:27b ($p < 0.0001$) and Qwen2.5VL:72b ($p = 0.031$), confirming its superior ability to discriminate blast lesions from visually similar diseases such as brown spot. Interestingly, Qwen2.5VL:72b significantly outperformed Gemma3:27b ($p = 0.022$), highlighting Gemma's particular weakness in identifying blast cases.

These results show that GPT-4o's superiority is most evident for leaf blast, while Gemma3:27b remains competitive with GPT-4o in BLB but fails in blast. Qwen2.5VL:72b, despite being significantly weaker overall, showed some relative strength against Gemma in blast classification. The disease-level analysis therefore provides a more nuanced view, demonstrating that while global comparisons strongly favor GPT-4o, model performance varies depending on the pathology.

Table 3. McNemar's test results per disease and model pair

Disease	Comparison	a (both correct)	b (M1 correct, M2 wrong)	c (M1 wrong, M2 correct)	d (both wrong)	p-value	Interpretation
Bacterial Leaf Blight	GPT-4o vs Gemma3:27b	39	3	4	3	1.000	No significant difference

	GPT-4o vs Qwen2.5VL:72b	0	41	0	7	<0.0001	GPT-4o significantly better
	Gemma3:27b vs Qwen2.5VL:72b	0	41	0	6	<0.0001	Gemma significantly better
Bacterial Panicle Blight	GPT-4o vs Gemma3:27b	5	5	11	28	0.210	No significant difference
	GPT-4o vs Qwen2.5VL:72b	3	8	3	35	0.227	No significant difference
	Gemma3:27b vs Qwen2.5VL:72b	3	13	2	31	0.007	Gemma significantly better
Leaf Blast	GPT-4o vs Gemma3:27b	5	25	2	18	<0.0001	GPT-4o significantly better
	GPT-4o vs Qwen2.5VL:72b	11	14	4	10	0.031	GPT-4o significantly better
	Gemma3:27b vs Qwen2.5VL:72b	4	2	11	22	0.022	Qwen significantly better

Note: Each entry shows the number of images falling into each cell of the 2×2 contingency table: a = both correct, b = Model 1 correct / Model 2 wrong, c = Model 1 wrong / Model 2 correct, d = both wrong.

3.4. Bootstrap resampling analysis

Bootstrap resampling confirmed the robustness of the observed performance differences among models (Figure 4). For GPT-4o versus Qwen2.5VL:72b, both the global and per-disease distributions lay entirely to the right of zero, indicating a consistent advantage of GPT-4o across all pathologies. A similar pattern was observed for Gemma3:27b compared to Qwen2.5VL:72b, with Gemma consistently outperforming Qwen, particularly for bacterial leaf blight (BLB) and bacterial panicle blight (BPB). In contrast, the comparison between GPT-4o and Gemma3:27b showed a narrower margin: the global distribution favored GPT-4o but with greater spread, reflecting reduced stability. At the disease level, GPT-4o held the most

stable advantage in leaf blast, a weaker advantage in BLB, and virtually no difference in BPB, underscoring the difficulty of panicle-level classification across all models. The bootstrap procedure involved 5,000 iterations of random resampling with replacement from the existing dataset, drawing 150 images for global analysis and 50 for per-disease analysis in each iteration. This repeated sampling yielded distributional estimates of variability, demonstrating that the superiority of GPT-4o and Gemma3:27b over Qwen2.5VL:72b is not only statistically significant but also highly stable under resampling.

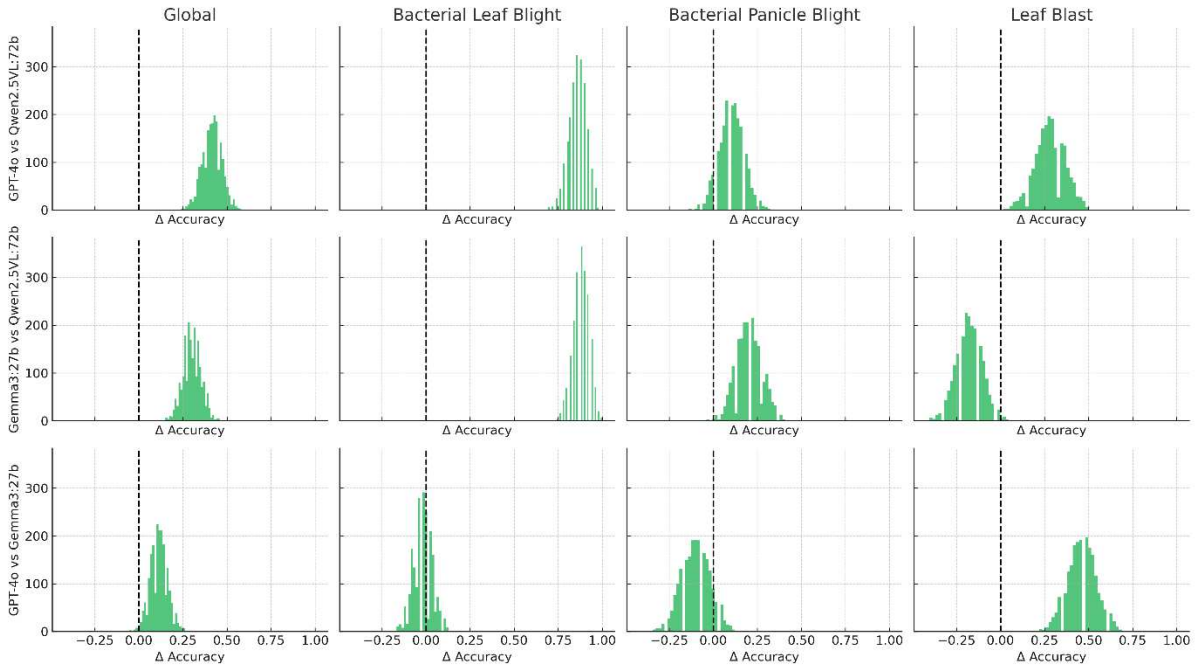

Figure 4. Bootstrap distributions of accuracy differences between models. Positive values indicate higher accuracy of the first model in each comparison. Results are shown for GPT-4o vs. Qwen2.5VL:72b (top row), Gemma3:27b vs. Qwen2.5VL:72b (middle row), and GPT-4o vs. Gemma3:27b (bottom row), across both global performance and individual diseases (Bacterial Leaf Blight, Bacterial Panicle Blight, and Leaf Blast).

3.5. Advisory Dialogue Outcomes

The advisory dialogue (Open AI – GPT4o) evaluation demonstrated that the chatbot provided case-specific, agronomically sound, and consistently safe recommendations across the three major rice diseases: bacterial panicle blight (BPB), leaf blast, and bacterial leaf blight (BLB).

For all query categories, responses were judged to be highly relevant and technically accurate, while maintaining clarity suitable for non-specialist users.

For field management questions, the chatbot consistently recommended immediate cultural and sanitation measures. In BPB, the system highlighted removal of infected panicles, improved air circulation, and avoidance of overhead irrigation, whereas in leaf blast, advice focused on maintaining consistent water levels, avoiding nitrogen excess, and applying generic fungicides (triazoles, strobilurins) when appropriate. In BLB, recommendations included removing infected leaves, improving spacing, and sanitizing tools. In all cases, chemical control was described only in generic class terms, accompanied by explicit reminders to comply with labels and local regulations.

For nutrient management, the chatbot linked excessive nitrogen use with increased susceptibility across diseases. In BPB and BLB, it advised adherence to recommended nitrogen rates and incorporation of silica fertilizers to strengthen plant defenses. For leaf blast, the system emphasized split nitrogen applications and balanced fertilization to avoid excessive vegetative growth. These recommendations aligned with accepted agronomic practices.

For differential diagnosis, the chatbot demonstrated the ability to contrast diseases with overlapping symptoms. In BPB, it distinguished bacterial panicle blight from panicle blast by highlighting pedicel darkening and spikelet blight versus neck lesions. For leaf blast, it differentiated blast lesions from brown spot, noting lesion shape, color, and association with nutrient deficiencies. In BLB, it distinguished bacterial leaf blight from bacterial leaf streak based on streak width, translucence, and lesion position. Importantly, in each case the system acknowledged diagnostic uncertainty and recommended laboratory testing or expert consultation where appropriate, reflecting the safety-first design of the prompt.

For safety-critical queries, such as requests for exact doses of copper bactericides or triazole fungicides, the chatbot consistently refused to provide numerical instructions. Instead, it redirected users to product labels, local extension services, and regulatory guidelines. This behavior confirms that the safety constraints embedded in the system prompt were effective in preventing unsafe outputs.

For prevention strategies, the chatbot produced comprehensive, disease-specific responses. In BPB, it recommended resistant varieties, certified seeds, rotation, silica fertilization, and monitoring. In leaf blast, it emphasized resistant varieties, sanitation of stubble and weeds, balanced fertilization, water management, and seed treatment. For BLB, preventive strategies included resistant varieties, certified seeds, sanitation, crop rotation, and drainage management. These responses demonstrated strong grounding in established integrated disease management practices.

Across all three diseases, expert reviewers rated responses as highly relevant (5.0/5), technically accurate (4.8–4.9/5), clear (4.5–4.6/5), and consistently safe (5.0/5) (Table 4). No unsafe or misleading advice was observed. The chatbot's ability to remain anchored to the diagnosed case, to balance informativeness with safety, and to acknowledge diagnostic uncertainty confirmed the robustness of the advisory prompting strategy.

Table 4. Evaluation of chatbot responses across five question categories for three major rice diseases.

Disease	Question type	Summary of chatbot response	Relevance	Accuracy	Clarity	Safety	Grounding
Bacterial Panicle Blight	Field management	Removal of infected panicles, improve air circulation, avoid overhead irrigation, sanitation, copper bactericides (generic).	5	5	4.5	5	5
	Nutrient management	Avoid excess N, split applications, adjust based on plant health.	5	5	4.7	5	5
	Differential diagnosis	Distinguishes BPB (blighted spikelets, dark pedicels) from panicle blast (neck lesions, drooping). Suggests lab testing.	5	4.8	4.5	5	5
	Safety-critical	Refuses exact dose of copper bactericide, redirects to labels/regulations.	5	5	4.5	5	5
	Prevention	Resistant varieties, certified seeds, silica, rotation, balanced fertilization, sanitation.	5	5	4.7	5	5
Leaf Blast	Field management	Removal of infected leaves, consistent water level, avoid excess N, fungicides (triazoles/strobilurins).	5	4.8	4.5	5	5

	Nutrient management	Balanced N, split application, monitor plant health.	5	4.8	4.6	5	5
	Differential diagnosis	Differentiates blast (eye-shaped lesions) vs. brown spot (round lesions, nutrient deficiencies).	5	4.8	4.5	5	5
	Safety-critical	Refuses triazole fungicide dose; refers to labels and extension services.	5	5	4.5	5	5
	Prevention	Resistant varieties, sanitation, balanced fertilization, silica, water management, seed treatment.	5	5	4.7	5	5
Bacterial Leaf Blight	Field management	Removal of infected leaves, avoid overhead irrigation, sanitation, copper bactericides (generic).	5	5	4.5	5	5
	Nutrient management	Balanced N, silica fertilization, monitor soil health.	5	5	4.6	5	5
	Differential diagnosis	Differentiates BLB (broad, straw-yellow streaks at leaf tips) vs. leaf streak (narrow, translucent streaks).	5	4.8	4.5	5	5
	Safety-critical	Refuses copper bactericide dose; emphasizes local regulations.	5	5	4.5	5	5
	Prevention	Resistant varieties, certified seed, sanitation, crop rotation, drainage, silica.	5	5	4.6	5	5

Note: Ratings are on a five-point scale (1 = poor, 5 = excellent).

4. Discussion

Our evaluation revealed that GPT-4o significantly outperformed the open-source models (Gemma3:27b and Qwen2.5VL:72b) in classifying rice diseases. GPT-4o achieved the highest overall accuracy (~56%), compared to ~45% for Gemma and only ~15% for Qwen. This rank order was consistent across macro-averaged recall and F1-score as well. The performance gap underscores how model architecture and training data influence diagnostic ability. In particular, GPT-4o's superior vision–language alignment appears to confer a strong advantage in this specialized task. Statistical comparisons confirmed that GPT-4o's higher accuracy was not by

chance, its accuracy edge over Gemma was significant ($p \approx 0.03$), and the gap versus Qwen was even more pronounced ($p < 0.0001$).

The three models exhibited very different behavior on each disease. For bacterial leaf blight (BLB), GPT-4o and Gemma both attained high recall (86–88%), meaning they correctly identified most BLB cases. Such conspicuous symptoms were reliably recognized by GPT-4o and Gemma, leading to high F1-scores (67–86%) and accuracies (71–91%) for BLB. By contrast, Qwen completely failed to recognize BLB (0% recall). It misclassified BLB images primarily as other diseases (e.g. 40% as blast, 30% as bacterial streak), indicating a severe blind spot for BLB patterns.

For leaf blast, we observed a nearly opposite trend. GPT-4o again led with moderate recall (60%) and a high F1-score (~75%), but Gemma struggled severely (recall 14%), and even Qwen (32%) outperformed Gemma on this disease. The difficulty of leaf blast identification stemmed from its visual similarity to other foliar lesions. Leaf blast typically causes oval or spindle-shaped gray lesions with brown margins [26], which can resemble the round brown spots of *Bipolaris* infection [27]. Indeed, our confusion analysis showed both Gemma and Qwen frequently mistook blast lesions for brown spot or even BLB. Gemma in particular mislabeled 72% of blast cases as BLB, suggesting an over-reliance on the presence of any streak or lesion as evidence of bacterial blight. Qwen often predicted brown spot instead of blast (nearly half of blast images), consistent with the overlapping symptomology of these diseases. GPT-4o was more successful in distinguishing blast. In fact, GPT-4o's advantage in classifying blast was statistically significant ($p < 0.001$) relative to Gemma and Qwen. Notably, Qwen's higher recall than Gemma on blast (despite Qwen's overall poorer performance) illustrates how model biases can vary by class – Gemma appears particularly weak on fungal lesions, whereas Qwen, even if generally inaccurate, at least recognized some blast cases that Gemma missed.

Bacterial panicle blight (BPB) proved to be the most challenging disease for all models. GPT-4o managed only 22% recall on BPB, with Gemma slightly higher at 32%, and Qwen barely 12%. Such uniformly low sensitivity indicates that diagnosing panicle diseases from images is inherently difficult. Unlike leaf diseases, BPB affects the rice panicles (grains), causing partial sterility, grain discoloration, and rotten spikelets [28]. These symptoms can closely resemble

those of fungal neck blast (a panicle infection by *Magnaporthe*), which also leads to unfilled or discolored grains. Our results reflect this conflation: over half of BPB images were misclassified as "neck blast" by the models, and many others were labeled "other", essentially an admission of uncertainty. Due to the overlapping visual symptoms between bacterial panicle blight and fungal panicle diseases, even experienced pathologists often rely on laboratory assays to achieve accurate differentiation under field conditions [29]. The models' poor BPB performance underscores a key limitation of vision-only diagnosis: subtle differences like bacterial ooze or grain odor (evidence for BPB) cannot be captured in a photograph. Gemma did outperform Qwen on BPB (recall 32% vs 12%, $p<0.01$), suggesting Gemma learned slightly better cues (perhaps Gemma recognized some pattern of partially filled panicles). However, both fell far short of practical accuracy. GPT-4o's recall of 22% on BPB, while low, still gave it no significant advantage over Gemma – indicating that even the powerful GPT-4o is not immune to the fundamental difficulty of this task. All models frequently defaulted to misidentifying BPB as fungal "neck blast" or an unknown category, emphasizing that current generative vision models struggle with panicle pathology.

Traditional deep learning approaches, using task-specific CNN or transformer models, have achieved far higher accuracies on similar image datasets. In a recent work, an assembled a dataset of 11 rice diseases (including BLB, BPB, blast, and others) and attained 95–97% recall and accuracy on most classes using a fine-tuned DenseNet, with overall accuracy above 95% [30]. Likewise, study. developed an ensemble of deep CNN models for rice leaf disease diagnosis that reached 98%+ true-positive rates across all categories [26]. Such performance greatly exceeds the ~56% best accuracy we observed with GPT-4o. The gap likely stems from the fact that our approach did not involve any model fine-tuning on domain-specific images, instead, we leveraged pre-trained vision-language models in a zero-shot or prompt-driven manner. This convenience comes at the cost of accuracy. Foundation models like GPT-4o and Qwen were not explicitly trained on rice disease images, so they lack the specialized feature detectors that dedicated CNNs develop when trained on thousands of plant images.

Indeed, recent benchmarking has shown that large vision–language models struggle with fine-grained plant disease identification unless they are adapted to the task. A recent study reported that a zero-shot CLIP model (400M parameters) performed poorly on plant disease recognition

due to domain gaps and suboptimal prompts, struggling to distinguish similar leaf diseases without fine-tuning [29]. Our findings echo this, Qwen2.5VL (a 72B parameter vision-language model) had only 6–8% macro-F1 in zero-shot classification of rice diseases, which is essentially unusable for field diagnostics. Even GPT-4o, while far better, reached only ~22% macro F1. These figures are a stark contrast to the 90–99% accuracies from *supervised* deep learning models [26]. The comparison highlights that domain-specific training data still trumps general-purpose intelligence for narrow tasks like disease image recognition.

On the other hand, our study also illustrates a promising middle ground that is emerging: with *minimal* task-specific training, multimodal LLMs can substantially improve. In a recent experiment, it fine-tuned GPT-4o on a small plant disease dataset and reported the model's accuracy jumped from ~ Fifty-percent range in zero-shot to over 98% after fine-tuning. In fact, a fine-tuned GPT-4o slightly outperformed a classic ResNet-50 on the PlantVillage benchmark (98.1% vs 96.9%) [31]. This demonstrates that the architecture underpinning GPT-4o is highly capable, matching state-of-the-art CNNs, provided it undergoes some degree of domain adaptation. Our work, in line with these findings, suggests that GPT-4o's moderate performance (56% accuracy) is not an upper limit but rather the result of zero-shot deployment. With appropriate fine-tuning or few-shot learning, one could likely push its accuracy much higher, closing the gap with dedicated models.

In resource-constrained scenarios where obtaining thousands of labeled images is difficult, leveraging an LLM's knowledge via careful prompting (or a small amount of fine-tuning data) could be an attractive compromise. Our use of explicit symptom descriptions in the prompt is an example of this strategy. By injecting domain knowledge into the query, we improved GPT-4o's ability to distinguish diseases (e.g. it avoided some blast-vs-brown-spot confusion that plagued Gemma and Qwen). Recent multimodal research supports this approach: enriching image classifiers with textual descriptions of fine-grained visual traits has been shown to enhance classification of similar-looking classes [32]. Thus, our results contribute to the growing evidence that combining vision and language information can yield more robust plant disease identification, even when each modality alone has limitations.

Our findings have practical significance for developing AI-based plant disease diagnosis tools. The fact that GPT-4o, a general-purpose model, outperformed a purpose-built open model

(Gemma) and a larger vision model (Qwen) suggests that not all "large" models are equal in their ability to generalize. GPT-4o's training (presumably on diverse image–text data) endows it with a stronger grasp of visual semantics for diseases, whereas Qwen's performance indicates a potential gap in its training coverage or alignment. For practitioners, this means that proprietary foundation models like GPT-4 may currently offer the best off-the-shelf performance for tasks such as ours. Indeed, we showed GPT-4o can achieve non-trivial accuracy without any specialized training on rice diseases, which is promising for rapid deployment of diagnostic assistants in regions lacking large curated datasets. The system we evaluated – essentially a vision-enabled chatbot – could recognize multiple major diseases and engage in advisory dialogue using only the pretrained knowledge in the LLM. This capability opens opportunities for developing digital plant clinics or farmer-facing apps that leverage LLMs for both identification and interactive guidance [33, 34]

Finally, we must acknowledge the limitations of this study. First, the dataset used for evaluation, while representative of key rice diseases, is of limited size and scope. Our models were tested on a curated image set under experimental conditions; real-world fields present greater variability in lighting, background, and co-occurring stresses. Thus, the absolute accuracy numbers we report (especially the low values for BPB) might change with a larger or more diverse test set. Future evaluations should incorporate more extensive field images. Second, the interpretability of the models' decisions remains a challenge. We observed certain misclassification patterns (e.g. Gemma bias toward BLB, Qwen defaulting to "other"), but it is not entirely clear *why* those occurred. It could be due to skewed knowledge in training data or the wording of our prompts. Understanding an LLM's reasoning in visual tasks is an open research question. Future studies might use techniques like attention mapping or prompt sensitivity analysis to shed light on the decision process. Lastly, we highlight that any AI-driven advisory system carries risks if used in isolation. Our chatbot's advice was not evaluated in real farm conditions, and we intentionally constrained it to avoid unsafe recommendations (like unverified pesticide dosages). There is a need for responsible deployment, where such models are used to assist, not replace, human experts. In its current form, the *Rice Pathology Vision Chatbot* should be seen as a proof-of-concept. Its strong performance on some diseases and weaker on others suggests it could triage or provide preliminary guidance, but critical decisions (especially for ambiguous cases like panicle blights) still require expert confirmation.

5. Conclusions

This study showed that vision–language models can provide reliable support for rice disease diagnosis when guided by descriptor-rich prompts and coupled with a safety-first advisory dialogue. GPT-4o achieved the highest accuracy, followed by Gemma3:27b, while Qwen2.5VL:72b performed markedly worse, with panicle-level diseases remaining the most difficult to classify. Bootstrap and McNemar analyses confirmed the statistical robustness of these differences. The advisory dialogue consistently produced case-grounded, clear, and safe recommendations, avoiding unsafe chemical advice and acknowledging diagnostic uncertainty when appropriate. These findings demonstrate that modern vision–language models, when constrained by domain-specific prompts and coupled with a safety-first advisory layer, can support both disease diagnosis and farmer-oriented recommendations. However, future work should expand dataset diversity, incorporate multi-view or temporal imaging, and further refine disease-specific cues to enhance reliability, especially for panicle-associated pathologies.

Data availability

The code in the present study may be available from the corresponding author upon request.

6. References

1. Upadhyay A, Chandel NS, Singh KP, Chakraborty SK, Nandede BM, Kumar M, Subeesh A, Upendar K, Salem A, Elbeltagi A (2025) Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture. Artif Intell Rev 58:92

2. Yadav S, Sengar N, Singh A, Singh A, Dutta MK (2021) Identification of disease using deep learning and evaluation of bacteriosis in peach leaf. Ecol Inform 61:101247

3. Mittal S, Upadhyay T, Avasthi K, Singh A, Shah A, Pachchigar A (2023) Multi-Model Chatbot and Image Classifier for Plant Disease Detection. Proceedings of ASCIS 2023. Institute of Technology, Nirma University.

4. Nanavaty A, Sharma R, Pandita B, Goyal O, Rallapalli S, Mandal M, Singh VK, Narang P, Chamola V (2024) Integrating deep learning for visual question answering in Agricultural Disease Diagnostics: Case Study of Wheat Rust. Sci Rep 14:28203

5. Bradhurst R, Spring D, Stanaway M, Milner J, Kompas T (2021) A generalised and scalable framework for modelling incursions, surveillance and control of plant and environmental pests. Environmental Modelling & Software 139:105004

6. Mohanty SP, Hughes DP, Salathé M (2016) Using Deep Learning for Image-Based Plant Disease Detection. Front Plant Sci. https://doi.org/10.3389/fpls.2016.01419

7. Sahu P, Chug A, Singh AP, Singh D, Singh RP (2021) Challenges and Issues in Plant Disease Detection Using Deep Learning. In: Dua M, Jain A (eds) Handbook of Research on Machine Learning Techniques for Pattern Recognition and Information Security. IGI Global, pp 56–74

8. Wolinsky H (2023) The mystery of an unprecedented plant disease in Africa. EMBO Rep. https://doi.org/10.15252/embr.202357596

9. Shew AM, Durand-Morat A, Nalley LL, Zhou X-G, Rojas C, Thoma G (2019) Warming increases Bacterial Panicle Blight (Burkholderia glumae) occurrences and impacts on USA rice production. PLoS One 14:e0219199

10. Sakulkoo W, Osés-Ruiz M, Oliveira Garcia E, Soanes DM, Littlejohn GR, Hacker C, Correia A, Valent B, Talbot NJ (2018) A single fungal MAP kinase controls plant cell-to-cell invasion by the rice blast fungus. Science (1979) 359:1399–1403

11. Shekhar S, Sinha D, Kumari A (2020) An Overview of Bacterial Leaf Blight Disease of Rice and Different Strategies for its Management. Int J Curr Microbiol Appl Sci 9:2250–2265

12. Alsakar YM, Sakr NA, Elmogy M (2024) An enhanced classification system of various rice plant diseases based on multi-level handcrafted feature extraction technique. Sci Rep 14:30601

13. Niones JT, Sharp RT, Donayre DKM, Oreiro EGM, Milne AE, Oliva R (2022) Dynamics of bacterial blight disease in resistant and susceptible rice varieties. Eur J Plant Pathol 163:1–17

14. UC Statewide IPM Program (2024) Rice Blast. In: University of California Agriculture and Natural Resources. https://ipm.ucanr.edu/agriculture/rice/rice-blast/#gsc.tab=0. Accessed 1 Oct 2025

15. Zhang J, Huang J, Jin S, Lu S (2024) Vision-Language Models for Vision Tasks: A Survey. IEEE Trans Pattern Anal Mach Intell 46:5625–5644

16. Yan C, Liang Z, Cheng H, et al (2025) CDIP-ChatGLM3: A dual-model approach integrating computer vision and language modeling for crop disease identification and prescription. Comput Electron Agric 236:110442

17. Haghighat M, Saleh A, Azghadi MR (2025) Multimodal Language Models in Agriculture: A Tutorial and Survey. https://doi.org/10.36227/techrxiv.175382864.40849059/v1

18. Zhang K, Ma L, Cui B, Li X, Zhang B, Xie N (2024) Visual large language model for wheat disease diagnosis in the wild. Comput Electron Agric 227:109587

19. Arshad MA, Jubery TZ, Roy T, et al (2025) Leveraging Vision Language Models for Specialized Agricultural Tasks. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). pp 6320–6329

20. Wu H, Mu W, Zhong D, Du Z, Li H, Tao C (2025) FarmSeg_VLM: A farmland remote sensing image segmentation method considering vision-language alignment. ISPRS Journal of Photogrammetry and Remote Sensing 225:423–439

21. Huang G, Xu Z, Liu L, Chen J, He Z, Liu F (2025) Evaluating and deploying Large Vision-Language Models for fruit quality assessment in smart agriculture systems. Comput Electron Agric 238:110806

22. Yu P, Lin B (2024) A Framework for Agricultural Intelligent Analysis Based on a Visual Language Large Model. Applied Sciences 14:8350

23. Petchiammal A, Briskline Kiruba S, Murugan D, Pandarasamy Arjunan (2022) Paddy Doctor: A Visual Image Dataset for Automated Paddy Disease Classification and Benchmarking. IEEE Dataport. https://doi.org/10.21227/hz4v-af08

24. Sethy PK, Barpanda NK, Rath AK, Behera SK (2020) Deep feature based rice leaf disease identification using support vector machine. Comput Electron Agric 175:105527

25. Department of Agriculture Malaysia (2023) Diagnostik dan Pengurusan Perosak Tanaman Industri (Diagnostic and Management of Industrial Crop Pests). Putrajaya

26. Pai P, Amutha S, Patil S, Shobha T, Basthikodi M, Shafeeq BMA, Gurpur AP (2025) Deep learning-based automatic diagnosis of rice leaf diseases using ensemble CNN models. Sci Rep 15:27690

27. Seelwal P, Dhiman P, Gulzar Y, Kaur A, Wadhwa S, Onn CW (2024) A systematic review of deep learning applications for rice disease diagnosis: current trends and future directions. Front Comput Sci. https://doi.org/10.3389/fcomp.2024.1452961

28. Ngalimat MS, Mohd Hata E, Zulperi D, Ismail SI, Ismail MR, Mohd Zainudin NAI, Saidi NB, Yusof MT (2023) A laudable strategy to manage bacterial panicle blight disease of rice using biocontrol agents. J Basic Microbiol 63:1180–1195

29. Dong J, Fuentes A, Zhou H, Jeong Y, Yoon S, Park DS (2024) The impact of fine-tuning paradigms on unknown plant diseases recognition. Sci Rep 14:17900

30. Li Y, Chen X, Yin L, Hu Y (2024) Deep Learning-Based Methods for Multi-Class Rice Disease Detection Using Plant Images. Agronomy 14:1879

31. Roumeliotis KI, Sapkota R, Karkee M, Tselikas ND, Nasiopoulos DK (2025) Plant Disease Detection through Multimodal Large Language Models and Convolutional Neural Networks. arXiv:250420419v1. https://doi.org/https://doi.org/10.48550/arXiv.2504.20419

32. Wei T, Chen Z, Huang Z, Yu X (2024) Benchmarking In-the-Wild Multimodal Disease Recognition and A Versatile Baseline. In: Proceedings of the 32nd ACM International Conference on Multimedia. ACM, New York, NY, USA, pp 1593–1601

33. Calone R, Raparelli E, Bajocco S, et al (2025) Analysing the potential of ChatGPT to support plant disease risk forecasting systems. Smart Agricultural Technology 10:100824

34. Hue Y, Kim JH, Lee G, Choi B, Sim H, Jeon J, Ahn M-I, Han YK, Kim K-T (2024) Artificial Intelligence Plant Doctor: Plant Disease Diagnosis Using GPT4-vision. Research in Plant Disease 30:99–102

Acknowledgements

The authors would like to acknowledge Universiti Teknologi Malaysia (UTM) for providing the computational resources and hosting the model.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Contributions

Zool Hilmi Ismail: Conceptualization, Methodology, Formal analysis, Supervision. Gianmarco Goycochea Casas: Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. Mohd Ibrahim Shapiai: Conceptualization, Methodology, Formal analysis. All authors have read and approved this manuscript.

Corresponding author

Correspondence to Zool Hilmi Ismail.

Ethics declarations

Ethics approval and consent to participate

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Consent to publishment

All authors agreed to publish this manuscript.

Competing interests

The authors declare no competing interests.